Explainable Anatomical Shape Analysis through Deep Hierarchical Generative Models

Carlo Biffi, Juan J. Cerrolaza, Giacomo Tarroni, Wenjia Bai, Antonio de Marvao, Ozan Oktay, Christian Ledig, Loic Le Folgoc, Konstantinos Kamnitsas, Georgia Doumou, Jinming Duan, Sanjay K. Prasad, Stuart A. Cook, Declan P. O'Regan, and Daniel Rueckert

***Abstract*—Quantification of anatomical shape changes currently relies on scalar global indexes which are largely insensitive to regional or asymmetric modifications. Accurate assessment of pathology-driven anatomical remodeling is a crucial step for the diagnosis and treatment of many conditions. Deep learning approaches have recently achieved wide success in the analysis of medical images, but they lack interpretability in the feature extraction and decision processes. In this work, we propose a new interpretable deep learning model for shape analysis. In particular, we exploit deep generative networks to model a population of anatomical segmentations through a hierarchy of conditional latent variables. At the highest level of this hierarchy, a two-dimensional latent space is simultaneously optimised to discriminate distinct clinical conditions, enabling the direct visualisation of the classification space. Moreover, the anatomical variability encoded by this discriminative latent space can be visualised in the segmentation space thanks to the generative properties of the model, making the classification task transparent. This approach yielded high accuracy in the categorisation of healthy and remodelled left ventricles when tested on unseen segmentations from our own multi-centre dataset as well as in an external validation set, and on hippocampi from healthy controls and patients with Alzheimer's disease when tested on ADNI data. More importantly, it enabled the visualisation in three-dimensions of both global and regional anatomical features which better discriminate between the conditions under exam. The proposed approach scales effectively to large populations, facilitating high-throughput analysis of normal anatomy and pathology in large-scale studies of volumetric imaging.**

***Index Terms*—Shape Analysis, Explainable Deep Learning, Generative Modeling, MRI.**

I. Introduction

THE quantification of anatomical changes and their relationship with disease is a fundamental task in medical image analysis, ultimately leading to new clinical insights and enhanced risk assessment and treatment. Recent improvements in the medical image analysis field have been characterised by an increase of large-scale population-based initiatives [1], [2], [3] together with development of automated segmentation pipelines of anatomical structures [4], [5], which recently achieved human-level performance [6]. In this context, the development of novel data-driven processing tools to enable quantitative assessment of the differences between normal anatomy and pathology has now received significant interest [7], [8], [9].

Alterations in shape and structure of an organ associated with an underlying pathology, here defined as pathological remodelling, are of particular interest for the classification and risk-stratification of patients. Hypertrophic cardiomyopathy (HCM) is a cardiac disease defined by the presence of left ventricular (LV) hypertrophy that cannot be solely explained by abnormal loading conditions [10]. In HCM, hypertrophy manifests in complex regional patterns not readily quantifiable using volumetric indices [11]. Similarly, atrophic changes in the hippocampus are considered as relevant biomarkers for the diagnosis and prediction of Alzheimer's disease (AD), and proved to differently affect distinct local areas of the hippocampal shape [12], [9]. For most human organs, the gold-standard imaging technique to assess structural shape changes is magnetic resonance (MR) which enables imaging at high-resolution and in three-dimensions (3D) [13], [5]. Despite the advances in MR imaging, classification and risk-stratification of patients still rely on scalar indexes describing pathological remodeling (e.g. left ventricular mass or hippocampal volume), which neglect regional or asymmetric effects that occur during pathology whose quantification could improve early detection and risk stratification [8], [13], [12], [9].

Machine learning approaches have achieved outstanding results in the medical image analysis domain, such as in the discrimination of physiological versus pathological hypertrophy patterns from multiple manually-derived cardiac indices [14], between patients with dilated cardiomyopathy patients and controls [15] and of patients with AD and mild cognitive impairment patients as well as healthy controls [16]. In particular, deep learning methods proved to be powerful features extractors for the classification of clinical conditions from medical images[17], [18]. Despite their tremendous success, however, a major drawback is their lack of interpretability, which currently hampers their translation to clinical practice. In fact, the physiological reason that drives the classification

The research was supported by grants from the British Heart Foundation (NH/17/1/32725, RE/13/4/30184), Academy of Medical Sciences (SGL015/1006) and the National Institute for Health Research Biomedical Research Centre based at Imperial College Healthcare NHS Trust and Imperial College London. *(Corresponding author: Carlo Biffi, e-mail: c.biffi15@imperial.ac.uk)*

C. Biffi, J. J. Cerrolaza, G. Tarroni, W. Bai, O. Oktay, C. Ledig, L. Le Folgoc, K. Kamnitsas, J. Duan and D. Rueckert are with the Department of Computing, Imperial College London. S. K. Prasad is with the National Heart and Lung Institute, Imperial College London. A. de Marvao, G. Doumou, D. P. O'Regan and S. A. Cook are with the MRC London Institute of Medical Sciences, Faculty of Medicine, Imperial College London.

result is often as important as the classification result itself [18].

In this work, we propose a new deep learning approach to learn a hierarchy of conditional latent variables that (1) models a population of anatomical segmentations of interest, (2) enables the classification of distinct clinical conditions by using the highest level of the hierarchy and (3) whose anatomical effect can be visualised and quantified in the original segmentation space. These contributions are achieved by specialising the highest level of a deep hierarchical generative model for the classification of distinct clinical conditions. As a consequence, thanks to the generative properties of the model, distinct segmentations corresponding to different values of the highest level can be generated, making the classification model interpretable. In addition, by constraining the highest level to be two-dimensional, the feature space in which the classification is performed can also be directly visualised. Therefore, our approach consists in an automated data-driven tool which enables the detailed analysis of the pathological remodelling patterns associated with a large number of clinical conditions.

II. Related Work

An autoencoder is a non-linear dimensionality reduction technique which learns a compact feature representation of the input data by encoding it into and decoding it from a low-dimensional feature vector. Deep autoencoder-based architectures have achieved wide success in computer vision applications as an extension of PCA-based approaches, including feature learning of 3D objects [19]. Autoencoder-based models have also been used to learn compact representations of medical images [17]. Relevant to this work, Oktay *et al.* [20] showed how autoencoder-derived features of LV segmentations outperform PCA features in the classification of healthy subjects versus dilated cardiomyopathy and HCM patients.

Deep generative models have demonstrated great performance in learning data distributions over a low-dimensional set of latent variables and in generating new unseen samples, which is not possible with standard autoencoder models. Within this class of models, variational autoencoder (VAE) models [21] learn a continuous latent representation by enforcing it to behave according to a predefined distribution. VAEs have been successful at learning the latent space representing deforming 3D shapes for a variety of applications, including shape space embedding and generation, outperforming state-of-the-art methods [22], [23]. In the medical imaging domain, VAEs have been exploited to approximate the distribution and likelihood of previously unseen MR images [24], to learn a low-dimensional manifold of 3D fetal skull segmentations [25] and to learn a low-dimensional probabilistic deformation model for cardiac image registration [26].

Hierarchical VAEs are a class of generative models that decompose the input data into a hierarchical representation [27], [28]. Although highly flexible, these models have been traditionally difficult to optimise, especially in the training of their higher levels, as often their lowest layer alone can contain enough information to reconstruct the data distribution, and the other levels are ignored. In this work, we focus on the ladder VAE (LVAE) framework [28], which was shown to be capable of learning a deeper and more distributed latent representation by combining the approximate likelihood and the data-driven prior latent distribution at each level of the generative model.

In hippocampus shape analysis, Shakeri *et al.* [29] employed a VAE model to learn a low-dimensional representation of co-registered hippocampus meshes, which was employed in conjuction with a multi-layered perceptron (MLP) to classify healthy subjects from AD patients. The network input consisted of mesh vertices coordinates, and the representation was learned through two fully connected layers. Similarly, in our preliminary work [30] we modified the 3D convolutional VAE framework in order to learn a low-dimensional latent representation of 3D LV segmentations, which was not only able to encode the 3D segmentations manifold, but also to discriminate different conditions by performing the classification task in the latent space. In the same work, we proposed a latent space navigation method to explore the anatomical variability encoded by the learned latent space. This consisted in iteratively modifying the latent representation of a segmentation obtained from an healthy subject along the direction that maximized its probability to be classified as pathological. By decoding the different latent representations in the original space of the segmentations, our technique allowed the visualisation of the anatomical changes caused by this transformation.

The following limitations characterize our preliminary work: 1) The learned VAE latent space not only encoded the factors of variation that most discriminate between classes, but also all the other factors of variation that regulate shape appearance. The latent space navigation was thus a necessary step to attempt the offline estimation of the variations linked to the pathological remodeling. In this work, we aim at automatically learning a latent space that encodes only these changes. 2) Our previous work required an additional offline dimensionality reduction technique to visualize in two dimensions the clustering obtained in the VAE latent space, which would however not reflect the real distribution of the shapes in the learned latent space. In this work, we aim at directly learning this two-dimensional latent space. 3) The latent space navigation method proposed in our previous work could only obtain subject-specific paths (with no obvious navigation stopping criteria). In this work, we aim at providing a means to extract the more clinically appealing population-based inferences.

In the later work of Bello *et al.* [31], a supervised denoising autoencoder was used to learn a latent code representation of right ventricular contraction patterns and, at the same time, to perform survival prediction. Not being a generative model, the effect of task-specific features learned by the proposed model could not be visualised, making the prediction task not explainable and population based inferences difficult to obtain. In addition, an additional offline dimensionality reduction step was also required to visualise in two-dimensions the distribution of different groups of subjects.

Contributions: In this paper, we aim to extend our preliminary work [30] on classification and visualisation of discriminative features by employing LVAEs, with the aim of assisting clinicians in quantifying the morphological changes related to disease, and in order to develop medical image classifiers that can visualise the morphological features driving the classification result. The main contributions of this work can be described as follows:

- We demonstrate that an interpretable classifier of anatomical shapes can be developed by performing a classification task of interest in the highest level of a LVAE model. In this way, the latent variables of this level automatically encode the most discriminative features for the task under exam, while the other subsequent levels model the remaining factors of anatomical variation in the data.
- We show that the LVAE highest latent space can be assumed to be two-dimensional so that the classification space can be directly visualised without further offline dimensionality reduction steps. Furthermore, we demonstrate how the anatomical variability encoded by this latent space can be visualised in the original space of the segmentations thanks to the generative properties of the model, enabling the visualisation of the anatomical effect of the most discriminative features between different conditions.
- We demonstrate how the proposed LVAE-based method achieves high classification accuracy of HCM versus healthy 3D LV segmentations and of AD versus healthy controls 3D hippocampal segmentations. More importantly, we show how the model captures and enables the easy visualisation of the most discriminative features between the conditions under exam. Finally, we show that the learned hierarchical representations provide higher reconstruction accuracy compared to single-latent-space VAEs.
- While hierarchical VAEs have been mainly evaluated on benchmark datasets, here we successfully apply them on two real-world 3D medical imaging datasets. We show insights on the model functioning and optimal training, and we make the implementation of proposed method publicly available[1].

III. Methods

This section is organised as follows. First, in subsection A and B, we summarise the theoretical foundations of the proposed method. Second, in subsection C, we describe our modifications to the original VAE and LVAE frameworks towards explainable shape analysis (graphical models in Fig. 1). Then, in subsection D, we describe the datasets used in this work for the classification of healthy subjects versus HCM patients and of heathy controls versus AD patients. Finally, in subsection E, we provide a detailed description of the LVAE models used in this work (model summary in Fig. 17 for the cardiac application).

[1]https://github.com/UK-Digital-Heart-Project/lvae_mlp
DOI 10.5281/zenodo.3247898

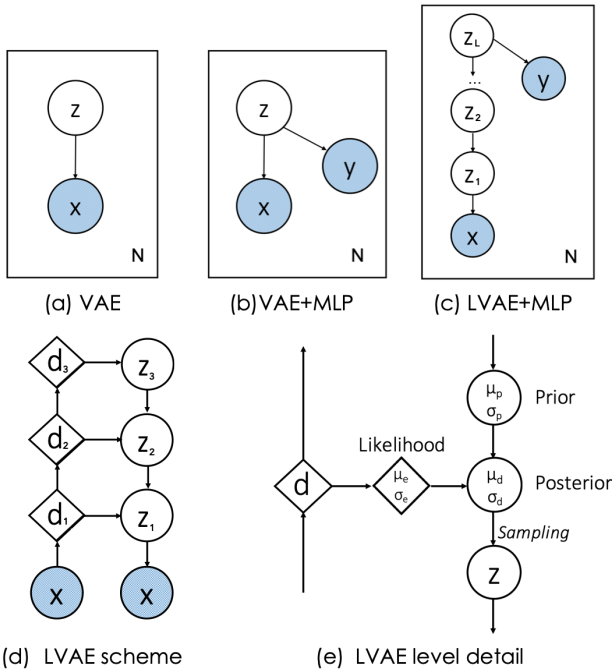

Fig. 1: Graphical models of a standard VAE (a), of our previously proposed method [30] (b) and the new LVAE-based approach (c). x represents and anatomical segmentation, y the disease class label and z the latent variables to learn. Schematic representation of a three-level LVAE (d) and of the flow of information (e). Circles represent stochastic variables, diamonds represent deterministic variables. Variables in light blue represent the inputs of the network.

A. Variational Autoencoder (VAE)

Given a training set of N anatomical segmentations $X = \{x_j,\ j = 1, ...N\}$ of a structure of interest from a population S, a VAE [21] is a probabilistic generative model that aims at learning the distribution $p_\theta(x)$ of the population of segmentations $x \in S$ under study. The distribution $p_\theta(x)$ is learned from the data by using a model of latent variables $z \in \mathcal{R}^p$, where $p \ll d$ and d is the number of pixels/voxels in a segmentation $x \in S$. The VAE graphical model is depicted in Fig. 1 (a) and the generative model is defined as

$$p_\theta(x) = \int_z p_\theta(x, z)dz = \int_z p_\theta(x|z)p_\theta(z)dz \qquad (1)$$

where $p_\theta(z)$ is the prior distribution over the variables z, $p_\theta(x|z)$ is the generative (or decoder) network and θ are the learnable parameters of the model. However, directly optimising $\log(p_\theta(x))$ for the N segmentations of the training set X is computationally infeasible, as it requires to compute the integral in Eq. 1 over all the z values. The VAE framework addresses this issue by introducing a variational distribution $q_\phi(z|x)$ to approximate the posterior distribution of the latent variables z, $p_\theta(z|x)$. After applying Bayes' rule and rearranging [32], the following equation can be derived

$$log(p(x)) - KL[q_\phi(z|x)||p_\theta(z|x)] = E_{q_\phi(z|x)}[\log(p_\theta(x|z))] - KL[q_\phi(z|x)||p_\theta(z)] \quad (2)$$

where KL is the Kullback-Leibler (KL) divergence. By assuming that the $q_\phi(z|x)$ is modeled with an high capacity function, the right-hand side of Eq. (2) becomes a lower bound for $\log(p_\theta(x))$ and can be optimized via stochastic gradient descent. The first term in the lower bound represents a reconstruction loss, i.e. how accurate is the generative model $p_\theta(x)$ in the reconstruction of the segmentation x from the latent space values z using the generative (or decoder) network $p_\theta(x|z)$. The second term is a regularization term that makes $q_\phi(z|x)$ match with its prior distribution $p_\theta(z)$ on the latent variables z.

B. Ladder Variational Autoencoder (LVAE)

A Ladder VAE (LVAE) [28] is a hierarchical latent variable model that employs a hierarchy of $i = 1, ..., L$ conditional latent variables in the generative model and it is schematised in Fig. 1 (d). The total prior distribution $p_\theta(z)$ of this model is factorised as:

$$p_\theta(z) = p_\theta(z_L) \prod_{i=1}^{L-1} p_\theta(z_i|z_{i+1}) \quad (3)$$

$$p_\theta(z_i|z_{i+1}) = \mathcal{N}(z_i|\mu_{p,i}(z_{i+1}), \sigma^2_{p,i}(z_{i+1})) \quad \forall i < L \quad (4)$$

$$p_\theta(z_L) = \mathcal{N}(z_L|0, 1) \quad (5)$$

where the highest latent space ($i = L$) has a prior distribution $p_\theta(z_L)$ which is typically assumed to be a Gaussian distribution with $\mu_{p,L} = 0$ and $\sigma^2_{p,L} = 1$ (Eq. 5), while the other levels in the hierarchy have their prior values of $\mu_{p,i}$ and $\sigma^2_{p,i}$ that conditionally depend on the upper levels of the ladder (Eq. 4).

The LVAE inference model also differs from a standard VAE. In particular, each layer i in the hierarchy of the latent variables is conditioned on the previous stochastic layers and the total inference model $q_\phi(z|x)$ is specified by the following fully factorised Gaussian distribution:

$$q_\phi(z|x) = q_\phi(z_1|x) \prod_{i=1}^{L-1} q_\phi(z_{i+1}|z_i) \quad (6)$$

$$q_\phi(z_i|\cdot) = \mathcal{N}(z_i|\mu_{d,i}, \sigma^2_{d,i}) \quad (7)$$

In contrast with standard hierarchical VAEs [27], where the inference $q_\phi(z|x)$ and prior distributions $p_\theta(z)$ are computed separately with no explicit sharing of information, the LVAE framework introduces a new inference mechanism. As shown in Fig. 1 (e), at each level i, an approximate likelihood estimation $\mu_{e,i}$ and $\sigma^2_{e,i}$ of its latent Gaussian distribution parameters is obtained from the encoder branch. This likelihood estimation is combined with the prior estimates $\mu_{p,i}$ and $\sigma^2_{p,i}$ obtained from the generative branch to produce a posterior estimation $\mu_{d,i}$ and $\sigma^2_{d,i}$ of the latent Gaussian distribution at that level i. In particular, this sharing mechanism between the inference (encoder) and generative (decoder) branches is performed at each level $i \neq L$ through a precision-weighted combination of the form:

$$\sigma^2_{d,i} = \frac{1}{\sigma^{-2}_{e,i} + \sigma^{-2}_{p,i}} \qquad \mu_{d,i} = \frac{\mu_{e,i}\sigma^{-2}_{e,i} + \mu_{p,i}\sigma^{-2}_{p,i}}{\sigma^{-2}_{e,i} + \sigma^{-2}_{p,i}} \quad (8)$$

while $\mu_{d,L} = \mu_{e,L}$ and $\sigma^2_{d,L} = \sigma^2_{e,L}$. This combination enables to build a data-dependent posterior distribution at each level, $\mathcal{N}(\mu_{d,i}, \sigma^2_{d,i})$, that is both a function of the values assumed in the higher levels of the generative model and of the inference information derived of the subsequent (lower) levels. The loss function of the LVAE is the same of a VAE (Eq. 2) with the only difference that the number of KL divergence terms is equal to the number of levels L in the ladder. These KL divergence terms force the learned prior and posterior distributions at each level to be as close as possible. The sharing of information between the encoder and decoder through Eq. 8 promotes the learning of a data-dependent prior distribution better suited for the dataset to be modelled. Moreover, this provides a better and more stable training procedure as the inference (encoder) branch iteratively corrects the generative distribution, instead of learning the posterior and prior values separately [28].

The full LVAE generative model has therefore the following formulation:

$$p_\theta(x) = \int_z p_\theta(x|z_1)\, p_\theta(z_L) \prod_{i=1}^{L-1} p_\theta(z_i|z_{i+1})\, dz \quad (9)$$

C. LVAE for Interpretable Shape Analysis

In our previous work [30], we proposed a modification of the standard VAE framework presented in Section III-A to include a classification network $p(y|z)$ able to predict the disease class label y associated with a segmentation x by using its latent representation z (the corresponding graphical model is shown in Fig. 1 (b)). In this work, we hypothesise that such modification can be extended to the LVAE framework by connecting a MLP $p(y|z_L)$, which classifies the disease status y of an input segmentation x, to the highest latent space z_L (graphical model in Fig. 1 (c)). By training the LVAE+MLP architecture end-to-end we aim at obtaining a very low-dimensional latent space z_L which encodes the most discriminative features for the classification task under study, while the other latent spaces will encode all the other factors of variation needed to reconstruct the input segmentations x. This has two main advantages: 1) template shapes for each disease class can be obtained by sampling from the learned distributions in a top-down fashion (starting from the highest level in the hierarchy $p(z_L|y)$ and subsequently from every prior $p_\theta(z_i|z_{i+1})$). The posterior $p(z_L|y)$ can be estimated by kernel density estimation and, since z_L is typically very low-dimensional, this estimation is straightforward; 2) if the latent space z_L is designed to be 2D or 3D, the distributions $p(z_L|y)$ in the classification space can be directly visualised without the need of further offline dimensionality reduction techniques required in previous works [30], [31].

D. Datasets

Cardiac Dataset A multi-centre cohort consisting of 686 HCM patients and 679 healthy volunteers was considered for this work. All subjects underwent cardiac phenotyping at a 1.5-T on Siemens (Erlangen, Germany) or Philips (Best, Netherlands) system using a standard cardiac MR protocol. HCM patients were confirmed with reference to established diagnostic criteria [13]. LV short-axis cine images were acquired with a balanced steady-state free-precession sequence. The end-diastolic (ED) and end-systolic (ES) phases were automatically segmented using a previously published and extensively validated cardiac multi-atlas segmentation framework [33]. As a first post-processing step, the obtained LV short-axis stack segmentations were upsampled using a multi-atlas label fusion approach. For each segmentation, twenty manually annotated high-resolution atlases at ED and ES were warped to the subject space using free-form non-rigid registration and fused with majority vote, leading to an upsampled high-resolution segmentation ($2mm$ x $2mm$ x $2mm$) [34]. In a second step, all segmentations were aligned onto the same reference space at ED by means of landmark-based and subsequent intensity-based rigid registration to remove pose variations. After extracting the LV myocardium label, each segmentation was cropped and padded to $[x = 80, y = 80, z = 80, t = 1]$ dimensions using a bounding box positioned at the centre of the LV ED myocardium. This latter operation guarantees shapes to maintain their alignment after cropping. Finally, all segmentations underwent manual quality control in order to discard scans with strong inter-slice motion or insufficient LV coverage, resulting in 436 HCM patients and 451 healthy volunteers that were used for the final analysis (population characteristics and standard CMR metrics are reported at Supplementary Data 6). As an additional external testing dataset, ED and ES segmentations from 20 healthy volunteers and 20 HCMs from the ACDC MICCAI17 challenge training dataset[2] were also used (after undergoing pre-processing using the same high-resolution upsampling pipeline explained above).

Brain Dataset A total of 726 3D left and right hippocampus segmentations of healthy controls (HC, $N = 404$, 202 males, median age 74.2 [min=59.8;max=89.6]) and Alzheimer's disease subjects (AD, $N = 322$, 177 males, median age 75.8 [min=55.1;max=91.4]) from a publicly available repository were analysed in this work [5]. The segmentations were obtained from baseline T1-weighted (T1w) MR brain images from the ADNI-1/-GO/-2 cohorts using a multi-atlas label propagation method with expectation-maximisation based refinement (MALPEM) [5]. Images were automatically segmented individually and no additional pre-processing was performed. All segmentations were rigidly registered to the MNI standard reference space using nearest neighbour interpolation. Shape-based interpolation was applied to upsample each segmentation to $0.75mm$ x $0.75mm$ x $0.75mm$ resolution. Finally, each segmentation was cropped and padded using a bounding box positioned at its centre to obtain 3D segmentations of dimension $[x = 60, y = 60, z = 60, t = 1]$ for both the left and right hippocampus. Moreover, a 3D high-resolution left and right hippocampus template segmentation was obtained by averaging the upsampled and rigidly registered healthy controls segmentations. By thresholding the template probabilistic segmentation, a template triangular mesh was extracted using marching cubes algorithm which will be used in this work for results visualisation.

[2]https://www.creatis.insa-lyon.fr/Challenge/acdc/

E. Application to Pathological Remodelling - LVAE+MLP model details

A detailed scheme of the three-level ($L = 3$) LVAE+MLP architecture employed in this work for the classification of HCM patients versus healthy subjects is summarised in Fig. 17, while the corresponding architecture for the classification of healthy controls versus AD patients is reported in Supplementary Materials 5. For the sake of display clarity the model scheme has been split into two rows: the encoder (inference) branch is shown at the top while the decoder (generative) branch is depicted at the bottom, and the two branches are connected by the latent space z_3. In the cardiac application, the input of the encoder branch are the 3D LV segmentations at ED and ES for each subject under study, which are presented as a two-channel input (top-left of Fig. 17). A 3D convolutional encoder compresses them into a 250-dimensional embedding through a series of 3D convolutional layers with stride 2. This embedding is used then as input of a deterministic inference network, which computes the likelihood estimates $\mu_{e,i}$ and $\sigma_{e,i}$ for each level i of the hierarchy of latent variables. These estimates are derived by manipulating the input through a series of fully connected layers (black arrows), which are all followed by batch normalisation and *elu* non-linearity with the only exception of the layers computing $\mu_{e,i}$ and $\sigma_{e,i}$. At the highest latent space ($i = 3$ in this case), a shallow MLP (2 layers) is attached to learn $p(y|z_3)$, i.e. to predict the class (HCM or healthy) label y corresponding to the input segmentation x by just using its latent variable values z_3. *ReLu* was used as non-linearity after the first layer. The latent variable values z_3 are sampled during training from $\mathcal{N}(\mu_{d,3}, \sigma^2_{d,3})$ where $\mu_{d,3} = \mu_{e,3}$ and $\sigma_{d,3} = \sigma_{e,3}$ and they are also the starting point of the generative process (bottom-right of Fig. 17). At each level i of the generative (decoder) network, the prior distribution terms are computed by modifying the values of the previous latent space z_{i+1} through a fully connected layer followed by batch normalization and *elu* non-linearity and by a second fully connected layer. These prior values are combined with $\mu_{e,i}$ and $\sigma_{e,i}$ through Eq. 8 to obtain the posterior estimates $\mu_{d,i}$ and $\sigma_{d,i}$ from which z_i is sampled. Finally, the value of z_1 is passed to a 3D convolutional decoder which aims to reconstruct the input segmentations x through a series of upsampling and convolutional layers. After every convolutional and upsampling layer used in the architecture *ReLu* was applied as non-linearity, except at the output of the network where *sigmoid* was applied. All the network weights were randomly initialised from a zero-mean Gaussian distribution ($\sigma = 0.02$).

The training loss function of the LVAE+MLP network is composed of three contributions: 1) two LV segmentation reconstruction accuracy terms at ED and ES as the overlap (Dice

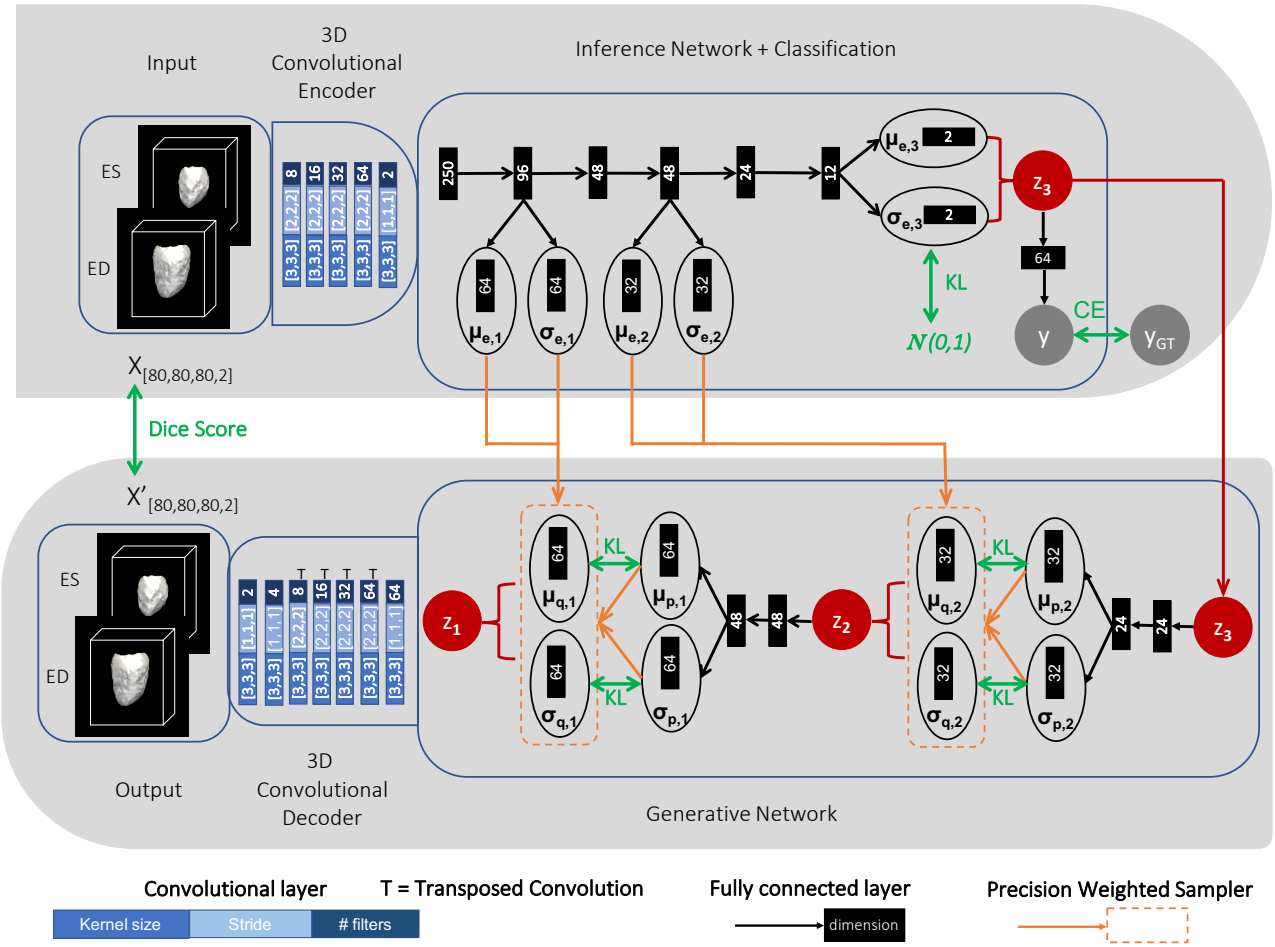

Fig. 2: Detailed scheme of the LVAE+MLP architecture adopted in this work for the cardiac application. Top: encoder model; Bottom: decoder model. At testing, segmentations class scores y are computed with $z_3 = \mu_{e,3}$. The green arrows indicate the loss function terms used to train the network.

score) between the input segmentation x and its reconstruction x'; 2) L KL divergence terms, penalising discrepancies between the learned prior and posterior distributions at each level and 3) a binary classification cross entropy (CE) term for the classification of healthy versus HCM segmentations. All the KL_i divergence terms except the one of the highest level ($i = 3$) were evaluated between the prior distribution $\mathcal{N}(\mu_{p,i}, \sigma^2_{p,i})$ and their posterior distribution $\mathcal{N}(\mu_{d,i}, \sigma^2_{d,i})$, while for the highest level the prior distribution was assumed to be a standard Gaussian $\mathcal{N}(0, 1)$. The total loss function is

$$\mathcal{L} = DSC_{ED} + DSC_{ES} + \gamma \left[\sum_{i=0}^{L} \alpha_i \, KL_i + \beta \, CE \right] \quad (10)$$

and depends on α_i, which weights the KL terms, on β, which weights the classification loss, and on γ, which is set to increase from 0 to 1 at the beginning of the training. This increase of γ is called deterministic warm-up and it has been commonly found useful in practice to converge to better local minima [28]. The weighting of the KL terms and the use of the Dice Score as a reconstruction metric lead to a different lower bound than standard VAE and LVAE. In the literature, it has been shown that the use of variants of the VAE lower-bound tend to favor better empirical results in various problems [35]. In this work, we adopted Dice score as reconstruction metric since it was successfully used in our previous work [30] and in related work [25] to achieve better reconstruction results on 3D anatomical segmentations.

At testing, a pair of ED and ES LV segmentations are reconstructed by starting from $z_3 = \mu_{d,3}$ and by assigning to z_2 and z_1 the values $\mu_{d,2}$ and $\mu_{d,1}$ computed from $z_3 = \mu_{d,3}$ and $z_2 = \mu_{d,2}$, i.e. no sampling is performed from the posterior distribution at each level. To interpret the anatomical information encoded by the highest latent space, at each level $i \neq 3$, the value of $\mu_{p,i}$ can be assigned to z_i instead of $\mu_{d,i}$ and the segmentations are reconstructed as explained above. In this way, by varying the values of z_3, a set of segmentations at ED and ES can be directly generated for each point in z_3, without using the inference information provided by $\mu_{e,i}$ and $\sigma_{e,i}$. This enables the visualisation of the anatomical information encoded by the highest latent space. Finally, in order to visualise the distribution of a set of segmentations under exam in the highest latent space, the $\mu_{e,3}$ values of each segmentation can be computed through the inference network and directly plotted in a 2D space.

Fig. 3: Latent space clusters in the highest latent space ($l = 3$) obtained by the proposed LVAE+MLP model on both the in-house training and testing datasets as well as on the ACDC dataset (entirely used as an additional testing dataset). Dimension 1 and 2 represent the two dimensions of $\mu_{e,3}$. On the left, long-axis sections of the reconstructed 3D segmentations at ED and ES obtained by sampling from three points in z_3 are shown.

IV. Results

A. Cardiac application

Model Training: Our in-house dataset of segmentations from healthy and HCM subjects was randomly divided into train, validation and test sets consisting of a total of 537 (276 from healthy volunteers, 261 from HCMs), 150 (75 from healthy volunteers, 75 from HCMs) and 200 (100 from healthy volunteers, 100 from HCMs) segmentations. We adopted a 3-level LVAE+MLP model (Fig. 17) since adding more levels did neither improve the reconstruction accuracy nor the classification accuracy in the clinical application under exam. The model was trained on a NVIDIA Tesla K80 GPU using Adam optimiser with learning rate equal to 10^{-4} and batch size of 16. For the first 40k iterations, data augmentation including rotations around the three standard axis with rotation angles randomly extracted from a Gaussian distribution $\mathcal{N}(0, 6°)$ was applied in order to take into account small mis-registrations. This helped the final model to achieve higher reconstruction accuracy, as it can be seen in the tables reported in the Supplementary Data 1. In the loss function (Eq. 10), the KL weights were fixed to $\alpha_1 = 0.02$, $\alpha_2 = 0.001$ and $\alpha_3 = 0.0001$ while γ was set to increase from 0 to 100 by steps of 0.5 every 4k iterations. The relative magnitude and ascending order of the KL weights α_i were chosen as they provided the best segmentation reconstruction results (i.e. higher Dice Score). In particular, our experiments showed that an ascending order of the weights improves both classification and reconstruction accuracy in contrast with models having all the weights α_i equal or in descending order (results are shown in Supplementary Data 2). This suggests the higher levels of a LVAE might be more difficult to train, and that a lower KL regularization term helps the training. The model produced similar results when varying these parameters within one order of magnitude, while a further increase in value reduced reconstruction accuracy and a further decrease resulted in model overfitting. The classification loss function weight β was instead set to 0.005: we observed that a higher value would have still produced a good model, but at the price of a more unstable training at the early stages. With regards to the number of layers and nodes adopted in the MLP, we have noticed that in general adopting a single fully connected layer poses a strong constraint on the latent space distribution, while using more than two causes overfitting. The increase of the classification and KL divergence weights during training through the γ parameter, known as deterministic warm-up [28], proved to be crucial to construct an expressive generative model (see in-depth analysis in Supplementary Data 1). After 220k iterations the training procedure was stopped as the increase of the KL divergence started to interfere with the decrease of the reconstruction and classification losses. In particular, this is due to the fact that in the highest latent space the KL divergence term tries to cluster all the data together, while the classification loss tries to separate the clusters. Hence the relative weight of β and α_3 needs to be tuned in order to obtain a good equilibrium.

VAE+MLP vs LVAE+MLP Reconstruction Accuracy				
	DSC_{ED}	DSC_{ES}	$H_{ED}[mm]$	$H_{ES}[mm]$
VAE+MLP train	0.81 ± 0.04	0.85 ± 0.04	6.30 ± 1.25	5.96 ± 1.20
LVAE+MLP train	0.85 ± 0.04	0.88 ± 0.03	5.70 ± 1.12	5.58 ± 1.00
VAE+MLP test	0.78 ± 0.04	0.83 ± 0.04	6.98 ± 1.65	6.75 ± 1.61
LVAE+MLP test	0.81 ± 0.04	0.85 ± 0.04	6.54 ± 1.62	6.40 ± 1.56

TABLE I: Cardiac. Dice score (DSC) and average 2D slice-by-slice Hausdorff distance (H) at ED and ES and their standard deviation for the proposed LVAE+MLP model and for the VAE+MLP model proposed in [30] on training and testing sets.

Classification and Reconstruction Results: All the 200 subjects in our testing dataset were correctly classified (100% sensitivity and specificity) by the trained prediction network. The same model also correctly classified 36 out of the 40 ACDC MICCAI 2017 segmentations (100% sensitivity and 80% specificity) without the need of any re-training procedure. Of note, 3 of the 4 misclassified segmentations suffered from a lack of coverage of the LV apex, which might be the cause for the error. The results obtained for the exemplar clinical

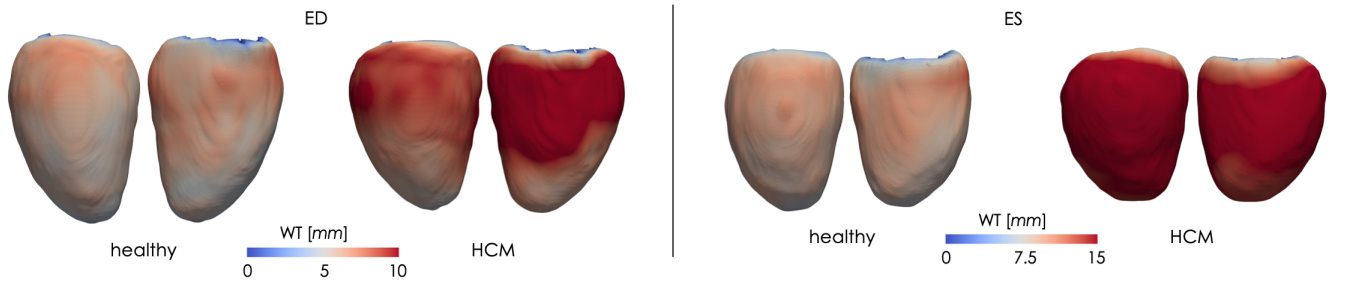

Fig. 4: Average healthy and HCM shapes at ED and ES sampled from the two clusters in the highest latent space of proposed LVAE+MLP model. The colormap encodes the vertex-wise wall thickness (WT), measured in mm.

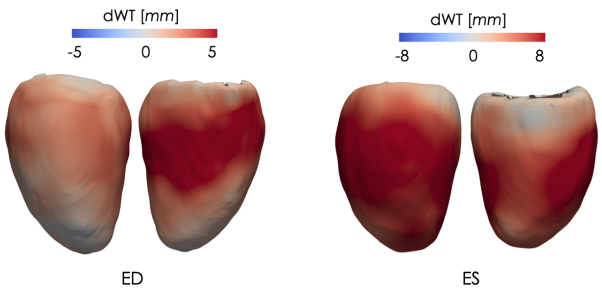

Fig. 5: Point-wise difference in wall thickness (dWT) at ED and ES between the healthy and the HCM average shapes of Fig. 4. Left - lateral wall; Right - septal wall.

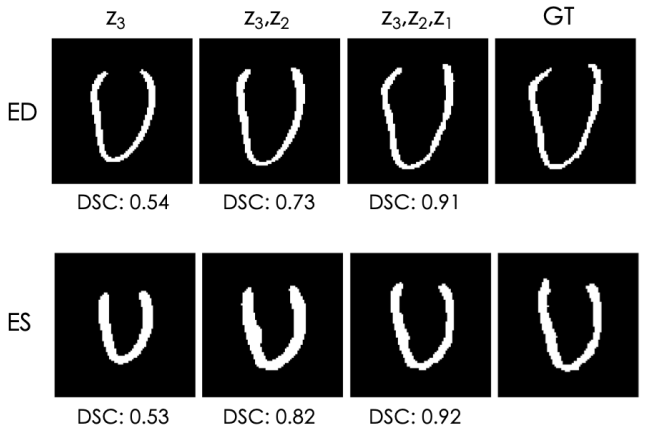

Fig. 6: Long-axis section of reconstructed segmentations at ED and ES by the LVAE+MLP model, using only z_3 information (first column) or also using the posterior information of the other latent spaces (z_2, z_1) . Last column: ground-truth (GT) segmentation. DSC = Dice Score between the segmentation at that column and the GT.

application are shown in Fig. 3, where two separated clusters of segmentations have been discovered both on the training and on the testing data. An analogous result was obtained in our previous work [30]: however, the previous version of the model required an additional dimensionality reduction step to visualise in 2D the obtained latent space of segmentations, while in the new proposed framework the highest latent space is 2D by design. Moreover, the new model achieved higher reconstruction accuracy than the previous model, as shown in Table VI, suggesting that a better generative model of shapes was learned. In particular, the table shows the reconstruction accuracy in terms of 3D Dice score and average 2D slice-by-slice Hausdorff distance between the 3D original and reconstructed segmentations on the testing and training datasets obtained by the proposed LVAE+MLP model and our previous VAE-based model (VAE+MLP) [30]. The VAE+MLP model was constructed with the same 3D convolutional encoder and decoder networks of the LVAE+MLP model and with a single latent space composed of 98 latent variables, which corresponds to the total number of latent variables adopted in the LVAE+MLP model (three levels of 64, 32 and 2 latent variables, respectively). As it can be noticed in the table, the obtained Dice score results at ES are better than at ED for all the models, while the Hausdorff results seem to follow instead an opposite trend. This is probably due to the fact that since the LV is more compact at ES, the Dice score might not be sensitive to small misalignment of the reconstructed shape, which are instead captured by the Hausdorff distance.

Visualisation of the latent spaces: Thanks to the properties of the proposed model, the anatomical information encoded by each latent space can be directly visualised, especially the information embedded in the highest level ($i = 3$), which encodes the most discriminative features for the classification of healthy and HCMs 3D LV segmentations. For the exemplar application under investigation, little intra-cluster variability between the shapes generated from the latent space z_3 was obtained, while much larger inter-cluster variability between the generated shapes was obtained. This can be seen on the left-side of Fig. 3 where we report a long-axis section of the 3D reconstructed segmentations at ED and ES at three points of the latent space z_3 (for a grid visualisation of the shapes encoded by this latent space please refer to Supplementary Data 3). Moreover, in Fig. 4 we show the obtained mean average shape for each cluster, represented as a triangular mesh with point-wise wall thickness (WT) values at vertex. This was obtained by sampling $N = 1000$ segmentations from each cluster in z_3 after estimating its probability density via kernel density estimation. Then, the obtained segmentations for each cluster were averaged to extract the corresponding average segmentation. Finally, a non-rigid transformation between the

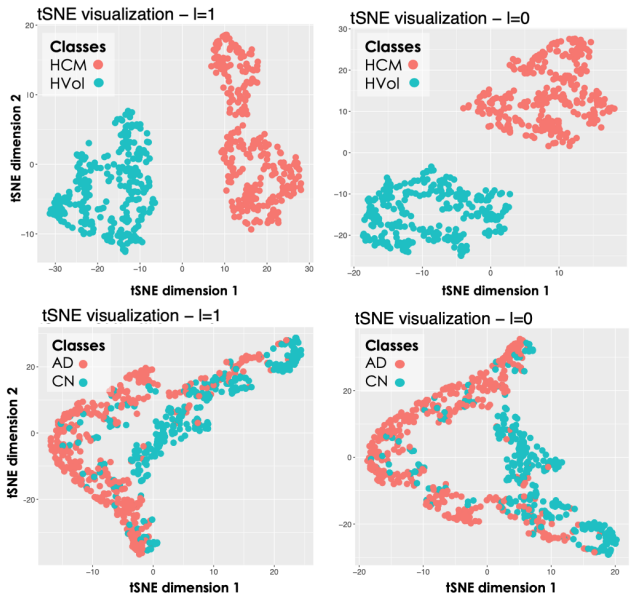

Fig. 7: tSNE visualisation of the latent spaces z_2 and z_1. Top: cardiac application. Bottom: brain application.

obtained average segmentation and a 3D high-resolution LV segmentation from the UK Digital Heart project[3] was computed, and the inverse of this transformation was applied to the corresponding 3D high-resolution LV segmentation to warp it to the cluster specific average segmentation. At each of the mesh vertices, WT was then computed as the perpendicular distance between the endocardial and epicardial wall. The results are presented in Fig. 4, where it can be noticed that the average HCM shape has higher WT than the corresponding healthy shape and it has a slightly reduced size. Fig. 5 instead reports the point-wise difference in WT between the HCM and the healthy shape, and it can be noticed that the most discriminative anatomical feature to classify an HCM shape consists in an increased WT in the septum, which is in agreement with the clinical literature [36]. Fig. 6 shows a long-axis section of the reconstructed segmentations at ED and ES from the LVAE+MLP model when only z_3 posterior information is used (first column) or when also the posterior information in the other levels (z_2, z_1) is exploited: the latent spaces z_2 and z_1 evidently encode different anatomical features that help to refine the structural information provided by z_3. Results for more subjects are reported in Supplementary Data 4. Finally, we applied the dimensionality reduction technique tSNE [37] to visualise in two dimensions the distributions of z_1 and z_2 latent spaces and we have found that the latent representations of the two classes of shapes are clustered also at both these levels (plots shown in Fig. 7). A possible explanation relies on the fact that the generative process is a conditional: if the data is clustered at the top of the hierarchy, it may be easier for the network to keep the clusters also in the subsequent levels.

[3]https://digital-heart.org/

VAE+MLP vs LVAE+MLP Reconstruction Accuracy				
	DSC_l	DSC_r	$H_r[mm]$	$H_l[mm]$
VAE+MLP train	0.81 ± 0.05	0.80 ± 0.05	3.35 ± 0.67	3.28 ± 0.69
LVAE+MLP train	0.85 ± 0.04	0.85 ± 0.03	3.05 ± 0.69	2.96 ± 0.66
VAE+MLP test	0.79 ± 0.05	0.79 ± 0.05	3.51 ± 0.64	3.49 ± 0.67
LVAE+MLP test	0.82 ± 0.03	0.82 ± 0.03	3.31 ± 0.68	3.23 ± 0.65

TABLE II: Brain. Dice score (DSC) and average 2D slice-by-slice Hausdorff distance (H) for the left (l) and right (r) hippocampus and their standard deviation for the proposed LVAE+MLP model and for the VAE+MLP model proposed in [30] on training and testing sets.

B. Brain application

Model Training: As an additional benchmark test, the LVAE+MLP model proposed in this work was also tested for the classification of healthy controls (HC) and patients with AD by using only 3D segmentations of the left and right hippocampus. Data was randomly assigned to train, validation and testing sets consisting of a total of 562 (322 HC, 240 AD), 64 (32 HC, 32 AD) and 100 (50 HC, 50 AD) segmentations respectively. A three level LVAE+MLP model was also adopted for this application (scheme in Supplementary Materials 5), since adding more levels did not improve classification or reconstruction accuracy. In the loss function (Eq. 10), the KL weights were fixed to $\alpha_1 = 0.03$, $\alpha_2 = 0.003$ and $\alpha_3 = 0.0003$, γ was set to increase from 0 to 100 by steps of 0.5 every 4k iterations and β was instead set to 0.005. The same augmentation strategy and the rationale for the selection of the hyperparameters in the previous experiment were adopted. The model training was stopped after 200k iterations.

Classification and reconstruction results: 84 out of 100 subjects were correctly classified by the training prediction network (78% sensitivity, 90% specificity). A VAE+MLP model with the same 3D convolutional encoder and decoder networks of the LVAE+MLP model, but with a single latent space of dimension 66 (equal to the total number of latent variables adopted in the LVAE+MLP model) was also trained. This model classified 81 out of 100 subjects correctly (74% sensitivity, 88% specificity) on the same training, testing and validation splits of the previous model. On the same dataset, an accuracy of 78% (75% sensitivity, specificity 80%) for the same classification task was obtained by using left and right hippocampus volume segmentations [5]. Compared to the VAE+MLP model, the LVAE+MLP model achieves higher reconstruction accuracy in terms of 3D Dice score and 2D slice-by-slice Hausdorff distance between the original segmentations and the reconstructed ones, these results are reported in Table II.

Visualisation of the latent spaces: Fig. 8 shows the distribution of the training and testing 3D hippocampus segmentations in the highest ($i = 3$) latent space for the trained LVAE+MLP model. It can be noticed how the healthy and pathological shapes are not as separated as in the previous application due to the more challenging nature of the new task. However, two clear clusters of healthy and AD shapes can still be identified. Fig. 8 also shows the left and right hippocampus

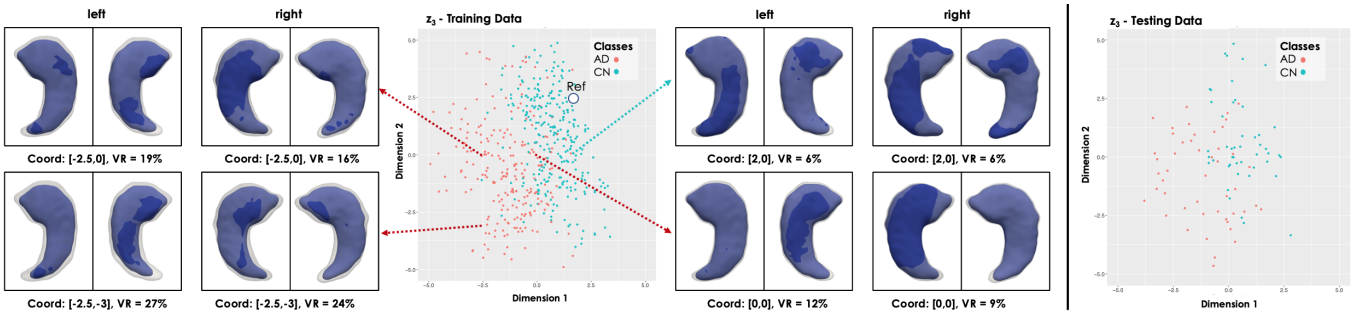

Fig. 8: Latent space clusters in the highest latent space ($l = 3$) obtained by the proposed LVAE+MLP model on the brain dataset. Left and right hippocampus shapes (in blue) at four points in the latent space have been reconstructed and showed together with a reference shape (in grey and opaque) sampled from the healthy control shapes (Ref, Coord: [2,2]). The first image is a view from the top, second image a view from the bottom.

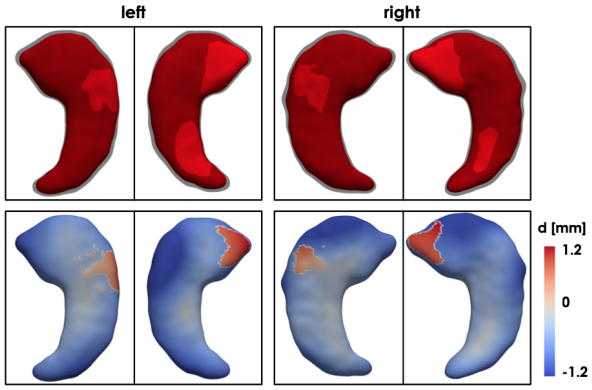

Fig. 9: First row: Average healthy (in grey and opaque) and AD (in red) left and right hippocampus shapes sampled from the two clusters in the highest latent space of proposed LVAE+MLP model. Second row: vertex-by-vertex L2 distance between the two mean shapes.

segmentations obtained by sampling from four distinct points of this latent space, which are showed together with a reference healthy shape sampled from a point in the healthy cluster (marked as Ref). For each reconstructed segmentation, the rate of hippocampal volume change (VR) with respect to the reference healthy shape was computed ($VR = \left|\frac{V-V_{ref}}{V_{ref}}\right| \times 100$). From the figure, it can be noticed how the AD shapes are characterized by decreased hippocampal volume, reduction that slighty but consistently affects more the left than the right hippocampus, in agreement with the previous findins on this data [5]. Moreover, a pattern in regional changes in volume can be identified: AD cases closer to the reference healthy shape show atrophy predominantly (if not only) in the tail of the hippocampus, while cases further away from the healthy class and deeper into the AD group show an atrophy pattern more spread throughout the whole hippocampal shape. In Fig. 9, we show the obtained average left and right hippocampus shapes from the healthy and AD distribution represented as triangular meshes. These meshes were obtained by sampling $N = 1000$ segmentations from the healthy and AD distributions in z_3 after estimating their probability density via kernel density estimation. Then, the obtained segmentations for each cluster were averaged to extract the corresponding cluster mean segmentation. Finally, the 3D template hippocampus segmentation was non-ridigly registered to each obtained cluster specific average segmentation and the estimated transformation was applied to the corresponding high-resolution mesh. In the first row of Fig. 9, it can be noticed how the reconstructed template AD segmentation (red) which is shown together with the HC segmentation (grey and opaque) is more atrophied and it is characterized by a bending of the head of both the left and right hippocampus. The second row displays the vertex-by-vertex L2 distance between the two mean shapes demonstrating a more pronounced regional atrophy in the hippocampal head consistent with the CA1 and subiculum regional athrophy already reported in the literature [9], [38]. The right hippocampus is characterized by a 13.5% decrease in volume between the healthy shape and the AD shape, while the decrease in volume for the left hippocampus is 14.6%. The volume ratio between the AD right and left hippocampus is 3.6% and 2.5% in the healthy mean shape. Finally, the plots resulting from the application of tSNE dimensionality reduction technique to the z_1 and z_2 training data values are shown at the bottom of Fig. 7.

V. Discussion

In this work, we presented a data-driven framework which learns to model a population of 3D anatomical segmentations through a hierarchy of conditional latent variables, encoding at the highest level of the hierarchy the most discriminative features to differentiate distinct clinical conditions. This is achieved by implementing and extending for the first time the LVAE framework to a real-world medical imaging application. In particular, by performing a classification task in the highest level of the LVAE hierarchy of latent variables, we can force this latent space to encode the most discriminative features for a clinical task under exam, while the other levels encode the other factors of anatomical variation needed to model the manifold of segmentations under analysis. Being a generative model, this framework provides the advantage of enabling the visualisation and quantification of the remodeling effect encoded by each latent space in the original segmentation

space. Hence, the anatomical differences used by the classifier to distinguish different conditions can be easily visualised and quantified by sampling from the highest level posterior distribution computed from a given database of shapes. Moreover, by designing this latent space to be two or three dimensional, no additional offline dimensionality reduction technique is required to visually assess the distribution of these shapes in the latent space. As a consequence, this method not only provides a deep learning classifier that uses a task-specific latent space in the discrimination of different clinical conditions, but more importantly enables the visualisation of the anatomical features encoded by this latent space, making the classification task transparent.

With the aim of assisting the clinicians in quantifying the morphological changes related to disease, we have applied the proposed framework for the automatic classification of heart and brain pathologies against healthy controls. In the reported cardiac application, the learned features achieved high accuracy in the discrimination of healthy subjects from HCM patients on our unseen testing dataset and on a second external testing dataset from the ACDC MICCAI 17 challenge. On the more challenging task of classification of healthy versus AD hippocampi, the model achieved better classification accuracy than using volumetric indices [5] and our previous method. Moreover, the visualisation of the features encoded in the highest level of the adopted LVAE+MLP model showed how the proposed model is able to provide the clinicians with a 3D visualisation of the most discriminative anatomical changes for the task under study, making the data-driven assessment of regional and asymmetric remodelling patterns characterizing a given clinical condition possible.

On both applications, we have also showed that the proposed LVAE+MLP model allows the construction of a better generative model in comparison to a VAE-based model with a single latent space [30]. To the best of our knowledge, this result confirms for the first time that hierarchical latent spaces provide a more accurate generative model on a real clinical dataset. Moreover, this work also gives insights on the functioning of these models on 3D anatomical segmentations, including how the different levels of latent variables encode different anatomical features, and how to optimally train this class of models for the reconstruction of these 3D anatomical segmentations.

While this work showed the potentialities of the proposed method on two common classification tasks, this method is domain-agnostic and could be applied to other classification problems where pathological remodelling is a predictor of a disease class label. However, further work is needed to explore the full potential of this approach, for instance in order to visualize the pathological remodeling of different disease subgroups characterized by different clinical endpoints. Of note, we expect that on very difficult tasks one or two more dimensions in the highest latent space might be needed, although further increasing the dimensionality will go against the rationale of the proposed approach. In fact, our aim is to encode the most discriminative anatomical information for the classification task under exam in the highest latent space, while the other latent spaces are intentionally left to model the remaining factors of variation. Interestingly, Fig. 7 shows that the shapes are clustered also in the other latent spaces, probably encoding additional variability of the disease groups not useful for the specific classification task. By specializing the classification task to more categories, we expect some information currently encoded in the other latent spaces to be moved and encoded in the highest one. For instance, studying multiple disease subgroups would enable a finer representation of the spectrum of remodeling patterns against which patients can be compared. Presently this was not achievable as the model has been optimized to discriminate only between healthy and diseased subjects, although a step in this direction was taken in Fig. 8, showing how different latent space points map to different hippocampal volume measures.

In comparison with the previous (generative) model [30] and Bello *et al.* [31] model, the proposed method requires tuning of a few additional hyperparameters, i.e. the number of adopted levels in the ladder and their weights importance in the model loss function. On the other hand, our approach is fully data-driven and it spares the need for further downstream dimensionality reduction and latent space navigation techniques, which would themselves require separate optimization and human intervention, potentially adding bias to the analysis. The proposed method also enables the derivation of population-based inferences (Fig. 4, 5 and 9), which could neither have been obtained from our previous model (due to the subject-specific nature of the latent-space navigation), nor from the one of Bello *et al.* (due to the non-generative nature of the model).

Another limitation shared by our previous and current approach is the fact that the input segmentations need to be rigidly registered to train the model. Future work should consider how to extend the proposed method to unregistered shapes, for example with the introduction of spatial transformer modules inside the architecture. In this work, as the output of the model is binary, Dice score was adopted as reconstruction metric. However, other alternatives exist, for example by modeling the model output with a Bernoulli distribution [32], and they will be investigated in future work. Finally, the prior distribution adopted in the highest latent space is a standard Gaussian distribution $\mathcal{N}(0, 1)$: future work could consider alternative prior distributions which could further favour the clustering of shapes. Even more interestingly, the interpretability and visualisation properties of the proposed method indicate that it could constitute an interesting tool for unsupervised clustering of shapes, for example by learning in the highest level discrete random variables.

VI. Conclusions

In recent years, the medical image analysis field has witnessed a marked increase both in the construction of large-scale population-based imaging databases and in the development of automated segmentation frameworks. As a consequence, the need for novel approaches to process and extract clinically relevant information from the collected data has greatly increased. In this work, we proposed a method for data-driven shape analysis which enables the classification

of different groups of clinical conditions through a very low-dimensional set of task-specific features. Moreover, this framework naturally enables the quantification and visualisation of the anatomical effects encoded by these features in the original space of the segmentations, making the classification task transparent. As a consequence, we believe that this method will be useful for the study of both normal anatomy and pathology in large-scale studies of volumetric imaging.

REFERENCES

[1] C. G. Fonseca *et al.*, "The Cardiac Atlas Project: an imaging database for computational modeling and statistical atlases of the heart," *Bioinformatics*, vol. 27, no. 16, pp. 2288–2295, 2011.

[2] R. Attar *et al.*, "Quantitative CMR Population imaging on 20,000 subjects of the UK Biobank imaging study: LV/RV quantification pipeline and its evaluation," *Medical Image Analysis*, 2019.

[3] S. G. Mueller *et al.*, "The alzheimer's disease neuroimaging initiative," *Neuroimaging Clinics*, vol. 15, no. 4, pp. 869–877, 2005.

[4] N. Nogovitsyn *et al.*, "Testing a deep convolutional neural network for automated hippocampus segmentation in a longitudinal sample of healthy participants," *NeuroImage*, vol. 197, pp. 589–597, 2019.

[5] C. Ledig, A. Schuh, R. Guerrero, R. A. Heckemann, and D. Rueckert, "Structural brain imaging in alzheimers disease and mild cognitive impairment: biomarker analysis and shared morphometry database," *Scientific reports*, vol. 8, no. 1, p. 11258, 2018.

[6] W. Bai *et al.*, "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks," *Journal of Cardiovascular Magnetic Resonance*, vol. 20, no. 1, p. 65, 2018.

[7] J. L. Bruse *et al.*, "Detecting clinically meaningful shape clusters in medical image data: metrics analysis for hierarchical clustering applied to healthy and pathological aortic arches," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 10, pp. 2373–2383, 2017.

[8] F. Triposkiadis *et al.*, "The continuous heart failure spectrum: moving beyond an ejection fraction classification," *European Heart Journal*, 2019.

[9] K. Shen *et al.*, "Detecting global and local hippocampal shape changes in alzheimer's disease using statistical shape models," *NeuroImage*, vol. 59, no. 3, pp. 2155–2166, 2012.

[10] C. W. Yancy *et al.*, "2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on practice guidelines," *Journal of the American College of Cardiology*, vol. 62, no. 16, pp. e147–e239, 2013.

[11] G. Captur *et al.*, "The embryological basis of subclinical hypertrophic cardiomyopathy," *Scientific Reports*, vol. 6, p. 27714, 2016.

[12] G. B. Frisoni *et al.*, "Mapping local hippocampal changes in alzheimer's disease and normal ageing with mri at 3 tesla," *Brain*, vol. 131, no. 12, pp. 3266–3276, 2008.

[13] M. Elliott Perry *et al.*, "ESC guidelines on diagnosis and management of hypertrophic cardiomyopathy: the task force for the diagnosis and management of hypertrophic cardiomyopathy of the european society of cardiology (ESC)," *European Heart Journal*, vol. 35, no. 39, pp. 2733–79, 2014.

[14] S. Narula *et al.*, "Machine-learning algorithms to automate morphological and functional assessments in 2D echocardiography," *Journal of the American College of Cardiology*, vol. 68, no. 21, pp. 2287–2295, 2016.

[15] E. Puyol-Antón *et al.*, "Regional multi-view learning for cardiac motion analysis: Application to identification of dilated cardiomyopathy patients," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 4, pp. 956–966, 2019.

[16] S. Basaia *et al.*, "Automated classification of alzheimer's disease and mild cognitive impairment using a single mri and deep neural networks," *NeuroImage: Clinical*, vol. 21, p. 101645, 2019.

[17] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[18] O. Bernard *et al.*, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[19] Y. Fang, J. Xie, G. Dai, M. Wang, F. Zhu, T. Xu, and E. Wong, "3D deep shape descriptor," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2319–2328.

[20] O. Oktay *et al.*, "Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation," *IEEE Transactions on Medical Imaging*, vol. 37, no. 2, pp. 384–395, 2018.

[21] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[22] C. Nash and C. K. Williams, "The shape variational autoencoder: A deep generative model of part-segmented 3D objects," in *Computer Graphics Forum*, vol. 36, no. 5. Wiley Online Library, 2017, pp. 1–12.

[23] Q. Tan, L. Gao, Y.-K. Lai, and S. Xia, "Variational autoencoders for deforming 3D mesh models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5841–5850.

[24] K. C. Tezcan, C. F. Baumgartner, R. Luechinger, K. P. Pruessmann, and E. Konukoglu, "MR image reconstruction using deep density priors," *IEEE Transactions on Medical Imaging*, 2018.

[25] J. J. Cerrolaza *et al.*, "3D Fetal Skull Reconstruction from 2DUS via Deep Conditional Generative Networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 383–391.

[26] J. Krebs, H. E. Delingette, B. Mailhé, N. Ayache, and T. Mansi, "Learning a probabilistic model for diffeomorphic registration," *IEEE Transactions on Medical Imaging*, 2019.

[27] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *arXiv preprint arXiv:1401.4082*, 2014.

[28] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," in *Advances in neural information processing systems*, 2016, pp. 3738–3746.

[29] M. Shakeri *et al.*, "Deep spectral-based shape features for alzheimers disease classification," in *International Workshop on Spectral and Shape Analysis in Medical Imaging*. Springer, 2016, pp. 15–24.

[30] C. Biffi *et al.*, "Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 464–471.

[31] G. A. Bello *et al.*, "Deep-learning cardiac motion analysis for human survival prediction," *Nature Machine Intelligence*, vol. 1, no. 2, p. 95, 2019.

[32] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[33] W. Bai *et al.*, "A bi-ventricular cardiac atlas built from 1000+ high resolution MR images of healthy subjects and an analysis of shape and motion," *Medical Image Analysis*, vol. 26, no. 1, pp. 133–145, 2015.

[34] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[35] T. Rainforth, A. Kosiorek, T. A. Le, C. Maddison, M. Igl, F. Wood, and Y. W. Teh, "Tighter variational bounds are not necessarily better," in *International Conference on Machine Learning*, 2018, pp. 4277–4285.

[36] M. Y. Desai, S. R. Ommen, W. J. McKenna, H. M. Lever, and P. M. Elliott, "Imaging phenotype versus genotype in hypertrophic cardiomyopathy," *Circulation: Cardiovascular Imaging*, vol. 4, no. 2, pp. 156–168, 2011.

[37] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[38] S. G. Mueller, N. Schuff, K. Yaffe, C. Madison, B. Miller, and M. W. Weiner, "Hippocampal atrophy patterns in mild cognitive impairment and alzheimer's disease," *Human brain mapping*, vol. 31, no. 9, pp. 1339–1347, 2010.

SUPPLEMENTARY DATA 1 - MODEL TRAINING - EFFECT OF DETERMINISTIC WARM-UP (DWU) AND DATA AUGMENTATION (DA)

Effect of DA and DWU					
Training					
	DSC_{ED}	DSC_{ES}	$H_{ED}[mm]$	$H_{ES}[mm]$	ACC [%]
None	0.75 ± 0.07	0.79 ± 0.05	7.30 ± 1.80	7.08 ± 1.68	51.40%
DA	0.77 ± 0.05	0.80 ± 0.05	6.94 ± 1.62	6.86 ± 1.53	51.40%
DWU	0.82 ± 0.05	0.86 ± 0.04	6.20 ± 1.23	5.93 ± 1.23	99%
DA+DWU	0.85 ± 0.04	0.88 ± 0.03	5.70 ± 1.12	5.58 ± 1.00	100%

TABLE III: Dice score (DSC) and average 2D slice-by-slice Hausdorff distance (H) at ED and ES and their standard error for the proposed LVAE+MLP model when Deterministic Warm-Up (DWU) and Data Augmentation (DA) are applied. ACC is the classification accuracy of the different models. Results obtained on the training dataset.

Effect of DA and DWU					
Testing					
	DSC_{ED}	DSC_{ES}	$H_{ED}[mm]$	$H_{ES}[mm]$	$ACC[\%]$
None	0.72 ± 0.07	0.76 ± 0.05	8.01 ± 1.99	7.53 ± 1.97	51.40%
DA	0.74 ± 0.06	0.78 ± 0.05	7.62 ± 1.86	7.31 ± 1.82	51.40%
DWU	0.79 ± 0.05	0.83 ± 0.04	6.91 ± 1.79	6.72 ± 1.68	99%
DA+DWU	0.81 ± 0.04	0.85 ± 0.04	6.54 ± 1.62	6.40 ± 1.56	100%

TABLE IV: Dice score (DSC) and average 2D slice-by-slice Hausdorff distance (H) at ED and ES and they standard error of the mean for the proposed LVAE+MLP model when Deterministic Warm-Up (DWU) and Data Augmentation (DA) are applied. ACC is the classification accuracy of the different models. Results obtained on the testing dataset.

SUPPLEMENTARY DATA 2 - MODEL TRAINING - EFFECT OF THE KL WEIGHTS

Effect of the KL weights					
Training					
[α_1,α_2,α_3]	DSC_{ED}	DSC_{ES}	$H_{ED}[mm]$	$H_{ES}[mm]$	$ACC[\%]$
$[10^{-4}, 2\ 10^{-4}, 10^{-3}]$	0.79 ± 0.05	0.80 ± 0.05	6.93 ± 1.62	6.88 ± 1.60	100%
$[10^{-4}, 10^{-4}, 10^{-4}]$	0.80 ± 0.05	0.83 ± 0.04	6.50 ± 1.41	6.48 ± 1.47	100%
$[10^{-3}, 2\ 10^{-4}, 10^{-4}]$	0.85 ± 0.04	0.88 ± 0.03	5.70 ± 1.12	5.58 ± 1.00	100%

TABLE V: Dice score (DSC), average 2D slice-by-slice Hausdorff distance (H) at ED and ES and they standard error of the mean together with classification accuracy (C) for the proposed LVAE+MLP model for different sets of the KL weights α_i in the training loss function. ACC is the classification accuracy of the different models. Results obtained on the training dataset.

Effect of the KL weights					
Testing					
[α_1,α_2,α_3]	DSC_{ED}	DSC_{ES}	H_{ED} [mm]	H_{ES} [mm]	$ACC[\%]$
$[10^{-4}, 2\ 10^{-4}, 10^{-3}]$	0.75 ± 0.06	0.78 ± 0.05	7.64 ± 1.72	7.37 ± 1.68	99%
$[10^{-4}, 10^{-4}, 10^{-4}]$	0.78 ± 0.05	0.80 ± 0.04	7.01 ± 1.53	6.94 ± 1.58	100%
$[10^{-3}, 2\ 10^{-4}, 10^{-4}]$	0.81 ± 0.04	0.85 ± 0.04	6.54 ± 1.62	6.40 ± 1.56	100%

TABLE VI: Testing data results. Dice score (DSC), average 2D slice-by-slice Hausdorff distance (H) at ED and ES and they standard error of the mean together with classification accuracy (C) for the proposed LVAE+MLP model for different sets of the KL weights α_i in the training loss function. ACC is the classification accuracy of the different models. Results obtained on the testing dataset.

SUPPLEMENTARY DATA 3 - CARDIAC APPLICATION - SEGMENTATIONS ENCODED BY z_3

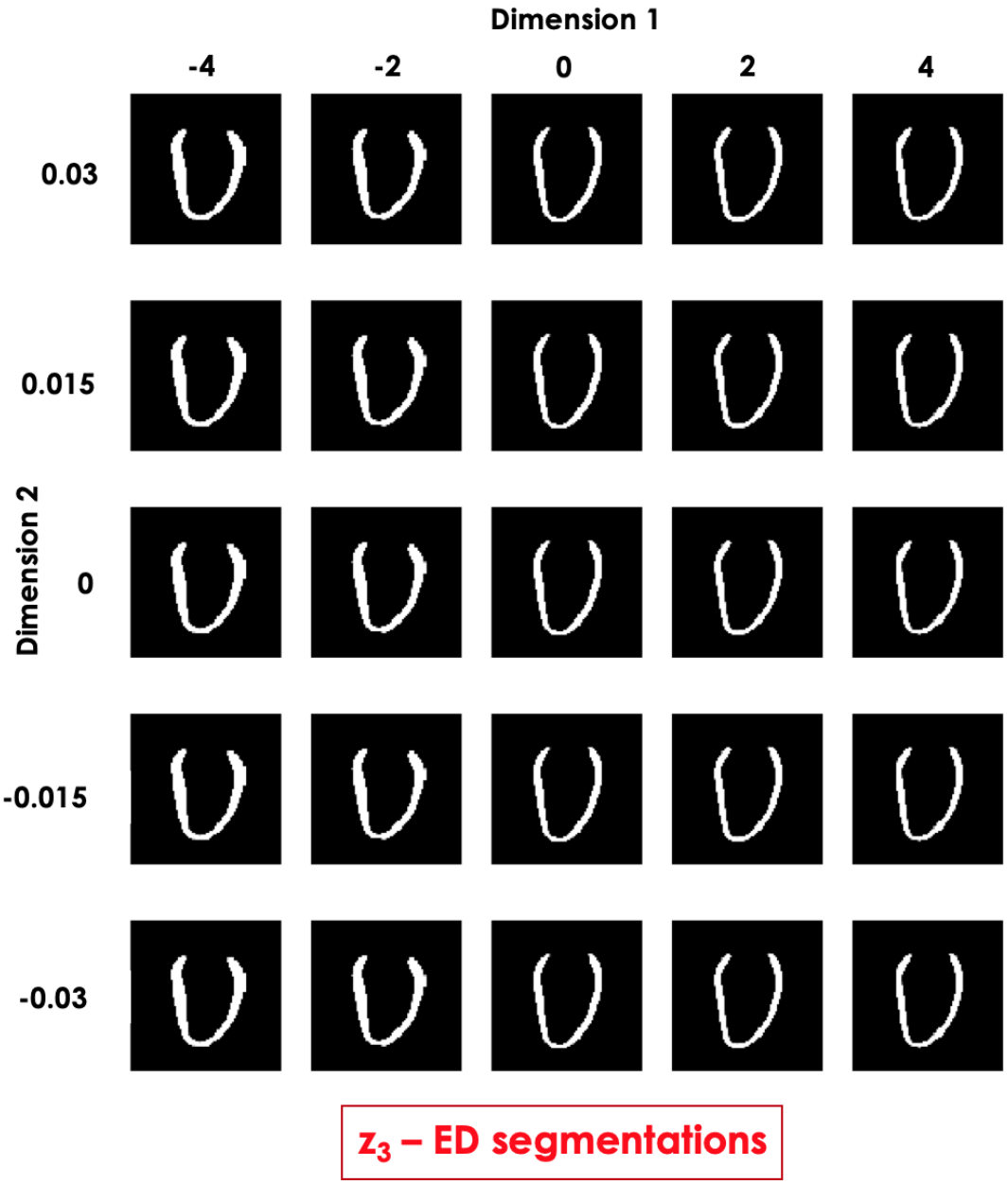

Fig. 10: Long-axis section of reconstructed segmentations at ED by the LVAE+MLP model by sampling from different points in z_3 and subsequently from the prior distribution of the latent variables z_2 and z_1.

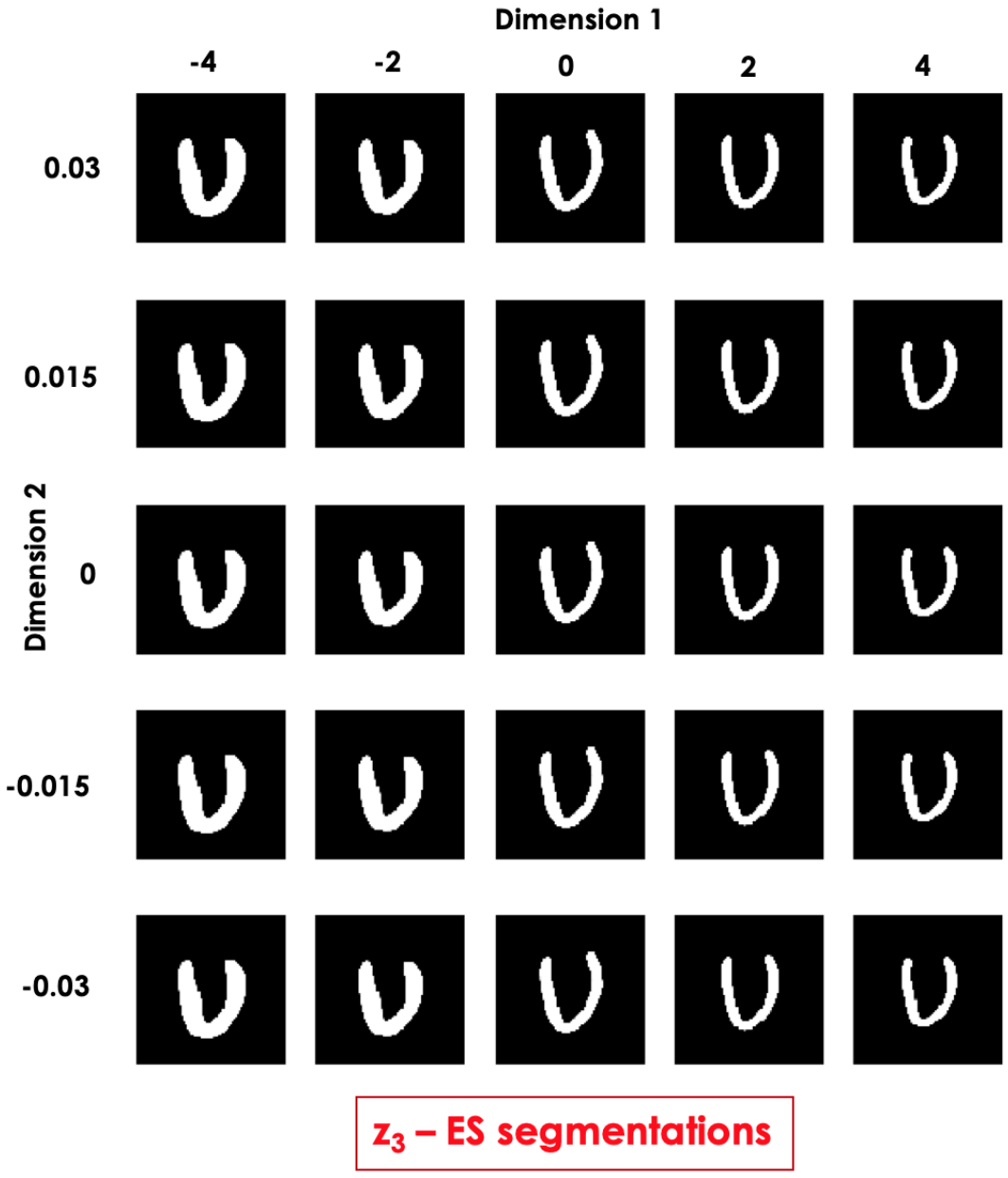

Fig. 11: Long-axis section of reconstructed segmentations at ES by the LVAE+MLP model by sampling from different points in z_3 and subsequently from the prior distribution of the latent variables z_2 and z_1.

SUPPLEMENTARY DATA 4 - CARDIAC APPLICATION - ADDITIONAL RECONSTRUCTION EXAMPLES

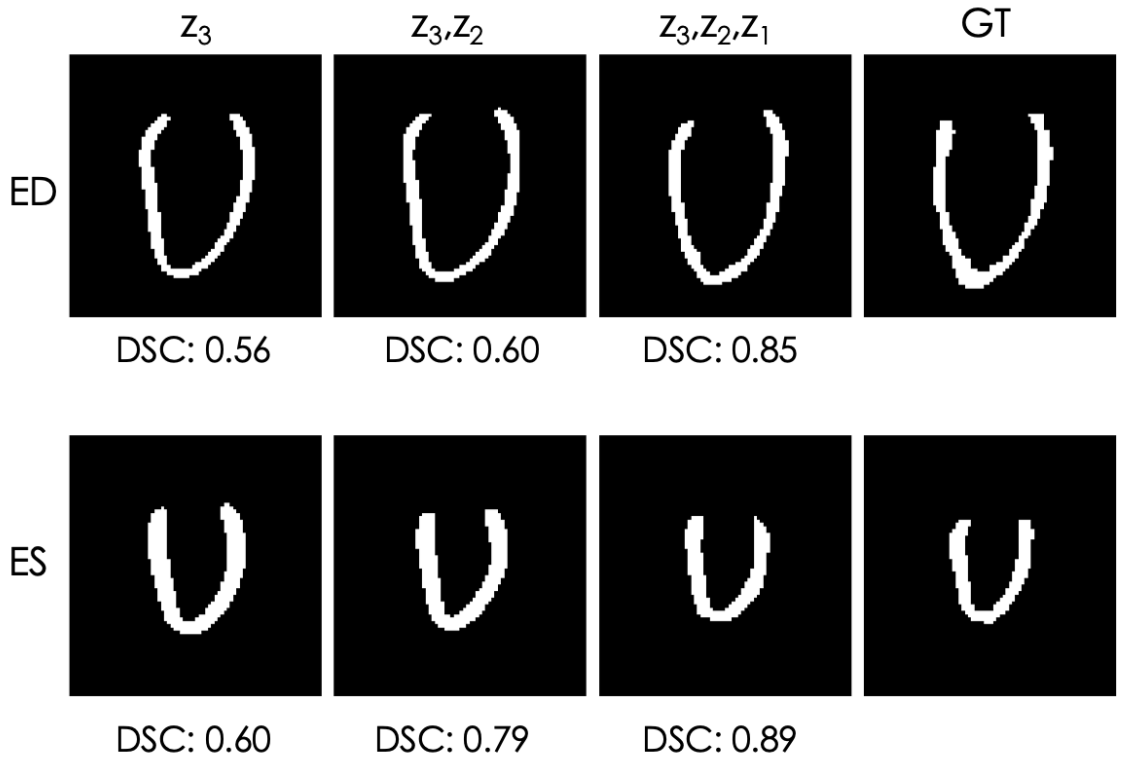

Fig. 12: Long-axis section of an healthy subject reconstructed segmentations at ED and ES by the LVAE+MLP model using only z_3 information (first column) or also using the posterior information of the other latent spaces (z_2, z_1) . Last column: ground-truth (GT) segmentation. DSC = Dice Score between the segmentation at that column and the GT.

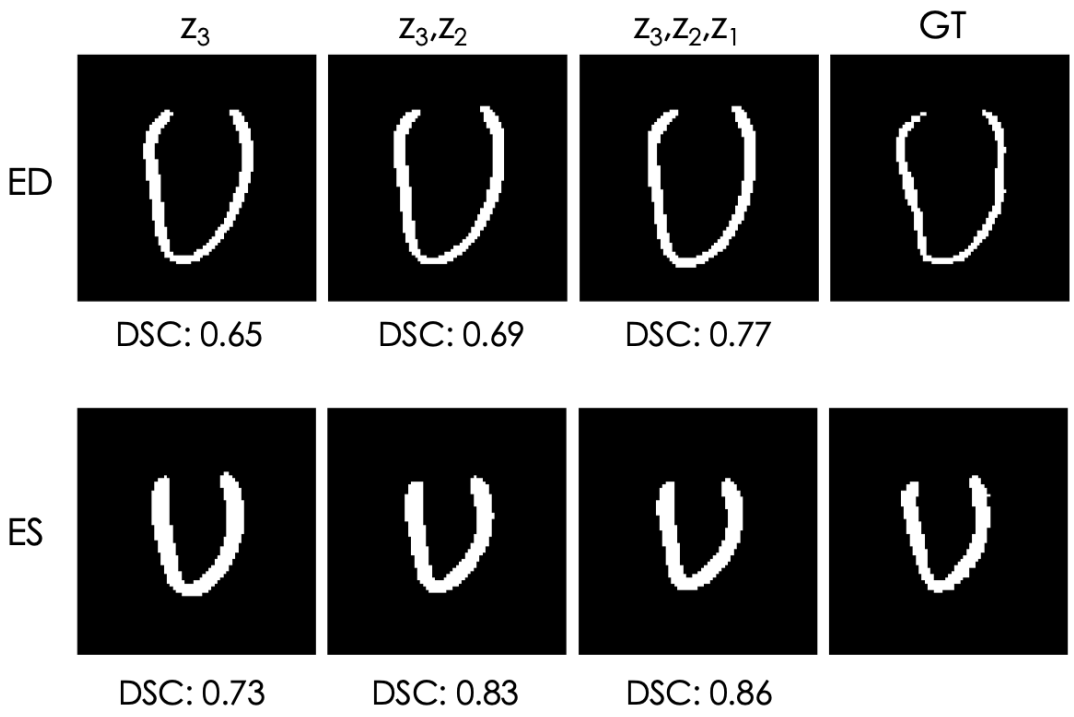

Fig. 13: Long-axis section of an healthy subject reconstructed segmentations at ED and ES by the LVAE+MLP model using only z_3 information (first column) or also using the posterior information of the other latent spaces (z_2, z_1) . Last column: ground-truth (GT) segmentation. DSC = Dice Score between the segmentation at that column and the GT.

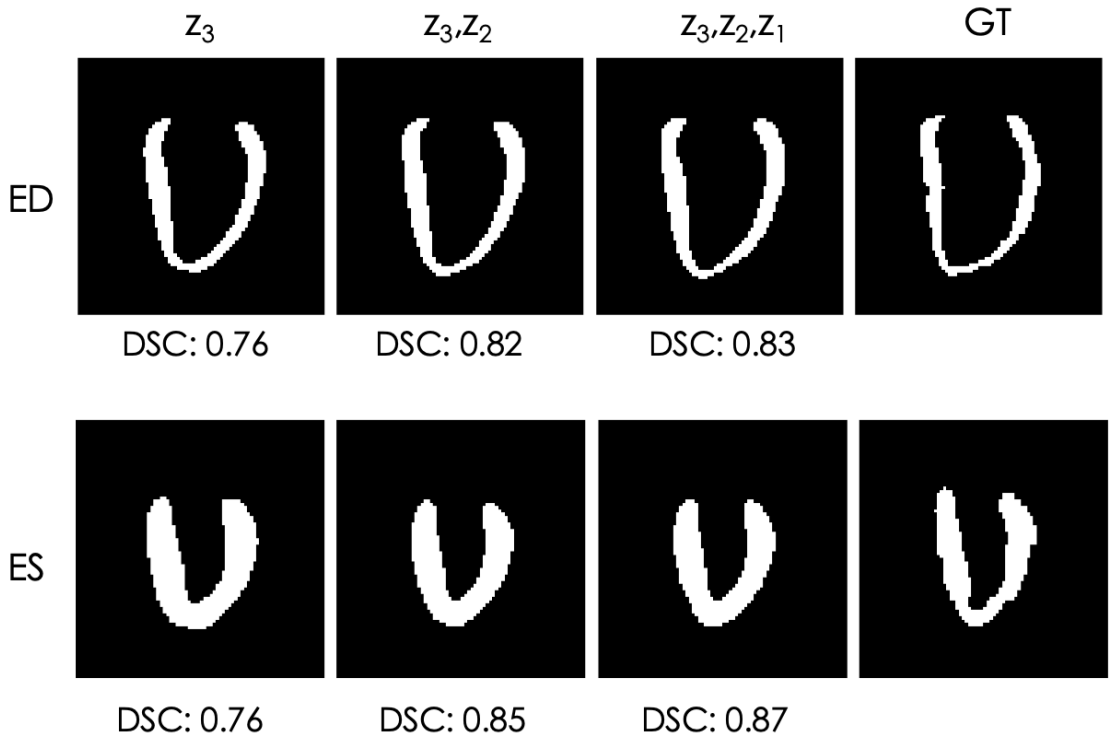

Fig. 14: Long-axis section of an HCM patient reconstructed segmentations at ED and ES by the LVAE+MLP model using only z_3 information (first column) or also using the posterior information of the other latent spaces (z_2, z_1) . Last column: ground-truth (GT) segmentation. DSC = Dice Score between the segmentation at that column and the GT.

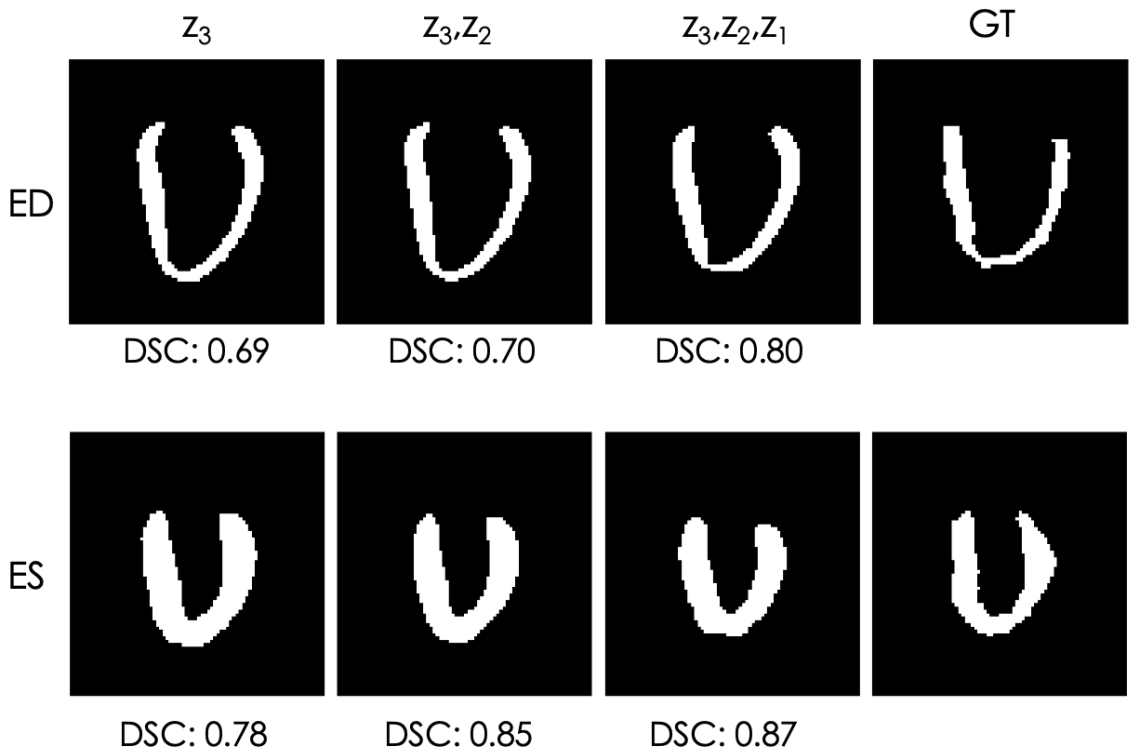

Fig. 15: Long-axis section of an HCM patient reconstructed segmentations at ED and ES by the LVAE+MLP model using only z_3 information (first column) or also using the posterior information of the other latent spaces (z_2, z_1) . Last column: ground-truth (GT) segmentation. DSC = Dice Score between the segmentation at that column and the GT.

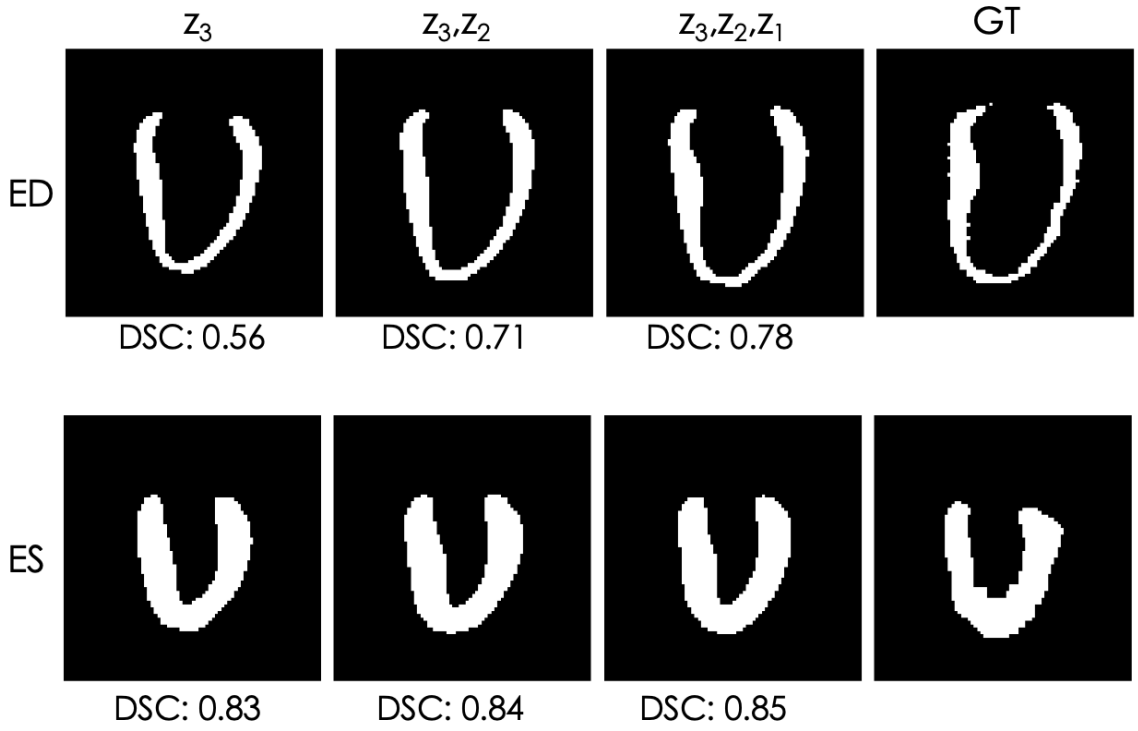

Fig. 16: Long-axis section of an HCM patient reconstructed segmentations at ED and ES by the LVAE+MLP model using only z_3 information (first column) or also using the posterior information of the other latent spaces (z_2, z_1) . Last column: ground-truth (GT) segmentation. DSC = Dice Score between the segmentation at that column and the GT.

SUPPLEMENTARY DATA 5 - BRAIN APPLICATION - LVAE+MLP ARCHICTURE

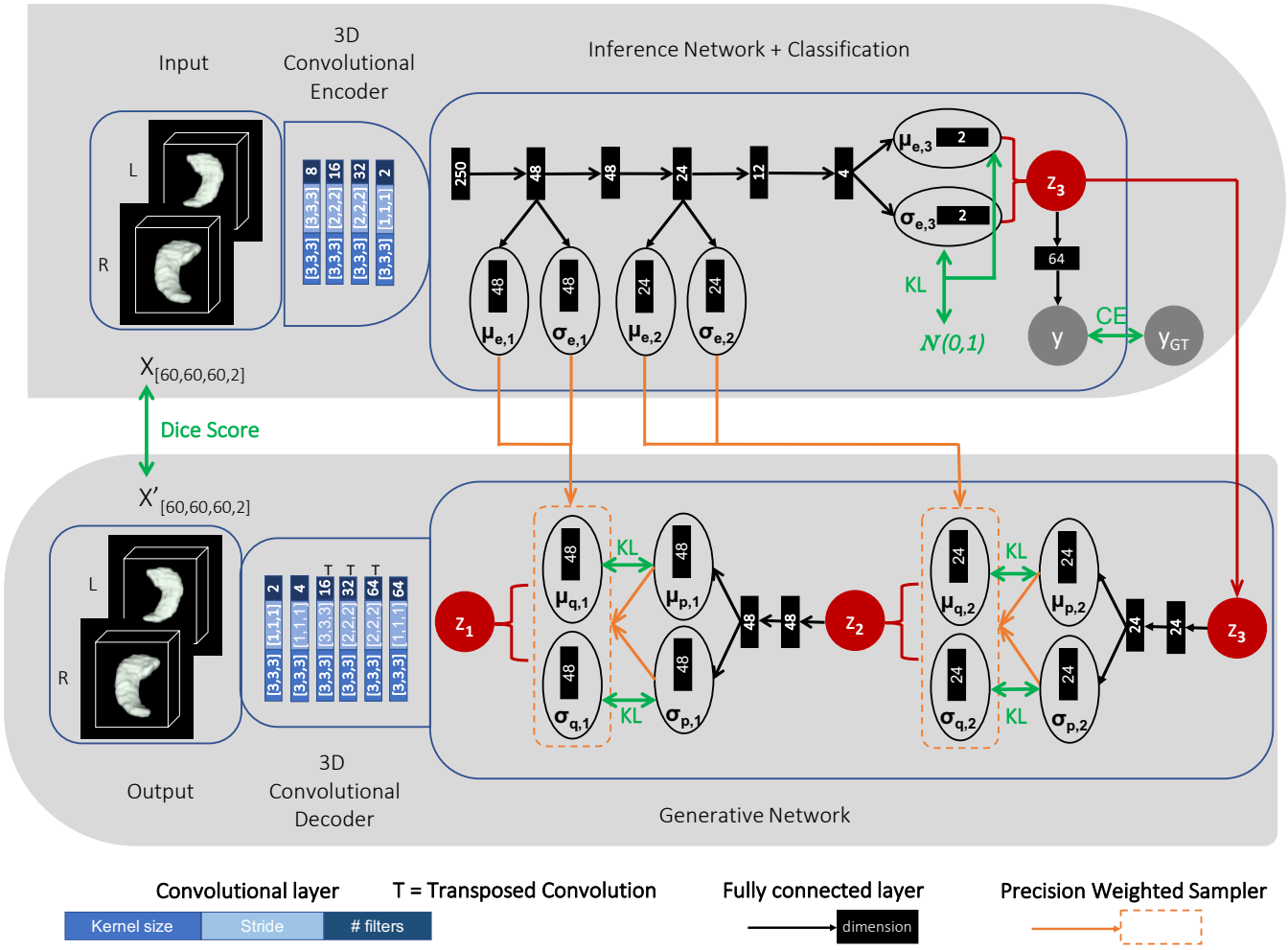

Fig. 17: Detailed scheme of the LVAE+MLP architecture adopted in this work for AD vs healthy controls experiments. Top: encoder model; Bottom: decoder model. The green arrows indicate the loss function terms used to train the network.

SUPPLEMENTARY DATA 6 - CARDIAC APPLICATION - STANDARD IMAGING MEASURES

	HCM		Hvol	
	Mean	SD	Mean	SD
Age at recruitment / first CMR	54.87	15.97	37.51	12.94
Females (%)	27.1		36.8	
BSA (m^3)	1.91	0.23	1.80	0.19
Left ventricular end-diastolic volume (mm^3)	134.77	36.35	141.77	30.52
Left ventricular end-systolic volume (mm^3)	35.28	17.43	48.53	14.51
Left ventricular ejection fraction (%)	74.45	9.74	66.12	5.04
Left ventricular mass (g)	182.00	64.84	109.84	32.29
Max wall thickness (mm)	18.33	4.85	7.37	3.52

TABLE VII: Table of population characteristics for the cardiac dataset. Information for 34 HCMs patients were not available. The total number of healthy volunters (HVols) subjects is 451, total number of HCMs is 402.